Speech Recognition
Lecture 9: Acoustic Models.

Mehryar Mohri
Courant Institute of Mathematical Sciences
mohri@cims.nyu.edu

Speech Recognition Components

- Acoustic and pronunciation model:

$$\Pr(o \mid w) = \sum_{d,c,p} \Pr(o \mid d) \Pr(d \mid c) \Pr(c \mid p) \Pr(p \mid w).$$

acoustic model

- $\Pr(o \mid d)$: observation seq. $\leftarrow$ distribution seq.
- $\Pr(d \mid c)$: distribution seq. $\leftarrow$ CD phone seq.
- $\Pr(c \mid p)$: CD phone seq. $\leftarrow$ phoneme seq.
- $\Pr(p \mid w)$: phoneme seq. $\leftarrow$ word seq.

- Language model: $\Pr(w)$, distribution over word seq.

Context-Dependent Phones

(Lee, 1990; Young et al., 1994)

- Idea:
  - phoneme pronunciation depends on environment (allophones, co-articulation).
  - model phone in context $\rightarrow$ better accuracy.
- Context-dependent rules:
  - Context-dependent units: $ae/b\_\_\_d \rightarrow ae_{b,d}$.
  - Allophonic rules: $t/V'\_\_\_V \rightarrow dx$.
  - Complex contexts: regular expressions.

Acoustic Models

- Critical component of a speech recognition system.
- Different types:
  - context-independent (CI) phones vs. context-dependent (CD) phones.
  - speaker-independent vs. speaker-dependent.
- Complex design and training techniques in large-vocabulary speech recognition.

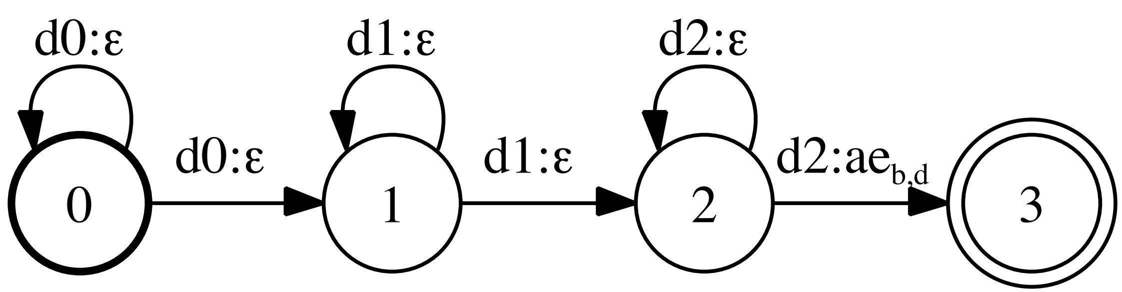
This Lecture

- Acoustic models
- Training algorithms

Continuous Speech Models

(Rabiner and Juang, 1993)

- Graph topology: 3-state HMM model: for each CD phone $ae_{b,d}$.



  - Interpretation: beginning, middle, and end of CD phone.

- Continuous case: transition weights based on distributions over feature vectors in $\mathbb{R}^N$, typically with $N = 39$.

Distributions

- Simple cases: e.g., single speaker, single Gaussian distribution

$$\mathcal{N}(x; \mu, \sigma) = \frac{1}{(2\pi)^{N/2}|\sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \sigma^{-1}(x-\mu)\right).$$

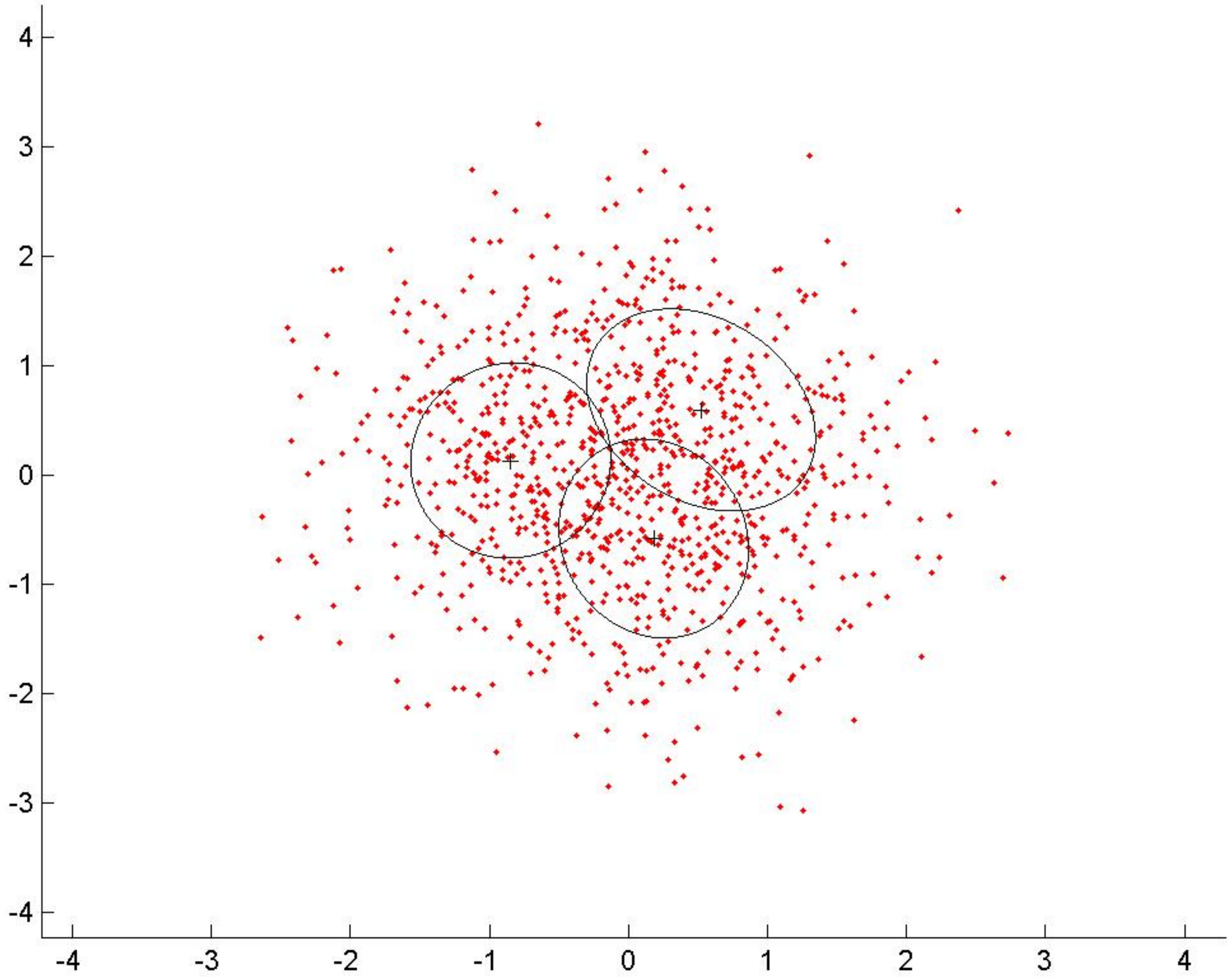
  - covariance matrix σ typically diagonal.

- General case: mixtures of Gaussians.

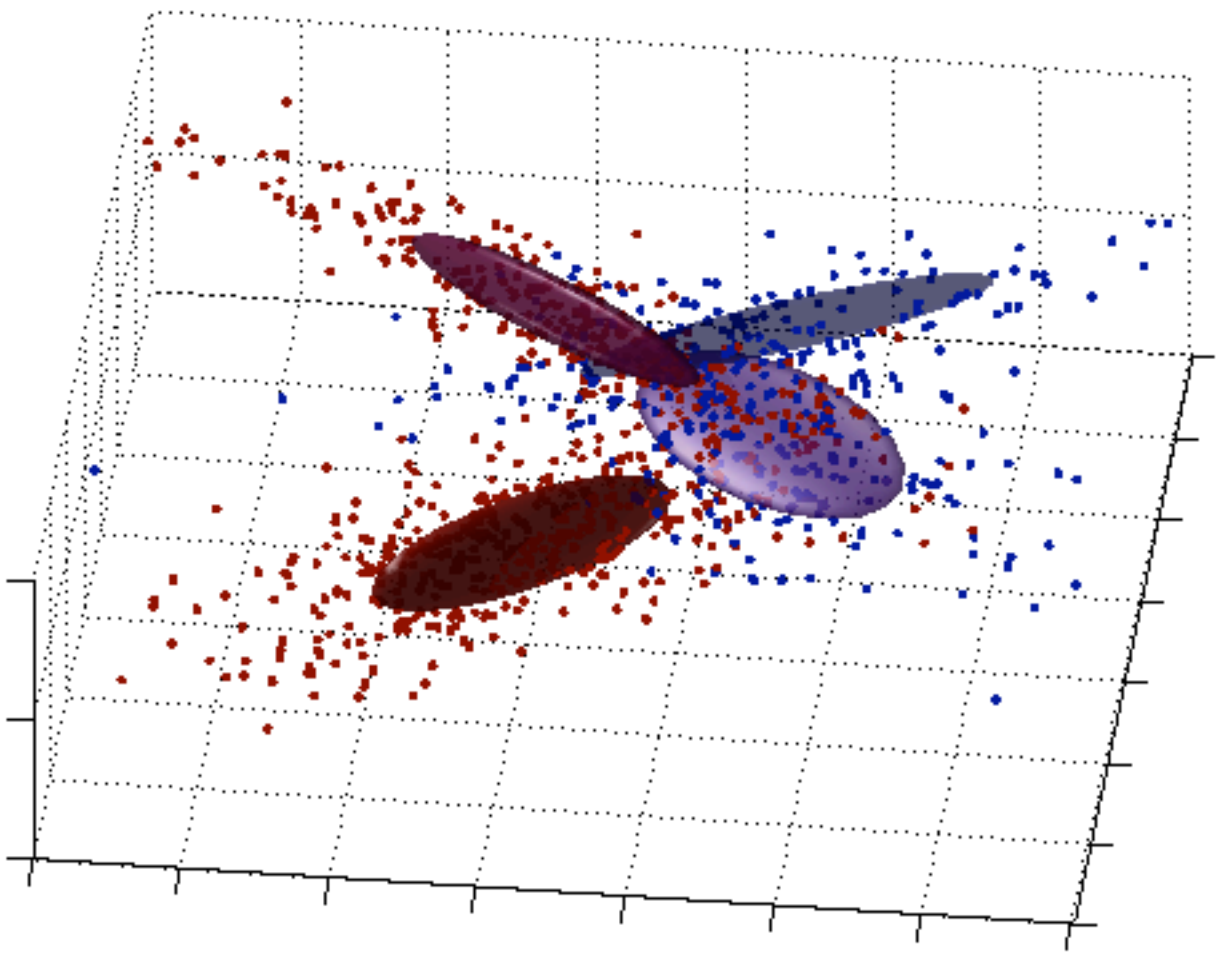
$$\sum_{k=1}^{M} \lambda_k \mathcal{N}(x; \mu_k, \sigma_k),$$

with $\lambda_i \geq 0$ and $\sum_{i=1}^{M} \lambda_i = 1$. Typically, $M = 16$.

GMMs - Illustration



GMMs - Illustration



Parameter Reduction

- Problem: too many parameters (> 200M).
  - large number of GMMs provides better modeling flexibility.
  - but requires much more training data.
- Solution: tying mixtures, i.e., equality constraints on distributions.
  - within the same HMM or distributions in different HMMs (similar CD phone transitions).
  - semi-continuous: same distrib. different mixtures.

Silence Model

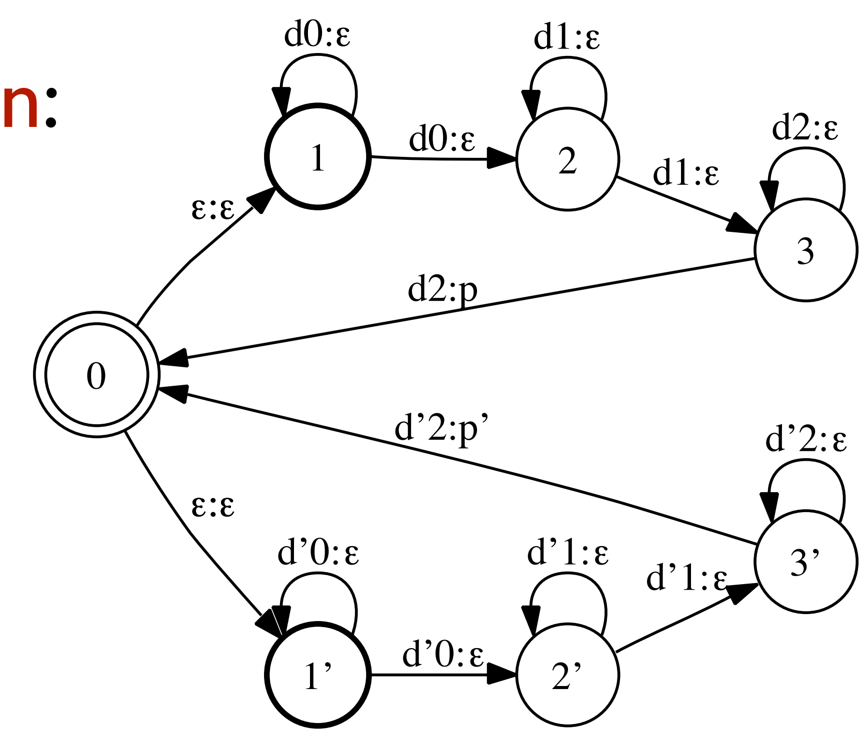
- Motivation: accounting for pause between words and sentences.
- Model:
  - optional pause symbol between words and at the beginning and end of utterances in language model.
  - specific silence acoustic model, which can be context-dependent or not.

Composite HMM model

- Composite model: obtained by taking the union and closure of all CD phone models.

$$\left(\sum_{p=1}^{P} H_i\right)^*.$$

- Illustration:



Tying can reduce the size.

This Lecture

- Acoustic models
- Training algorithms

Parameter Estimation

- Data: sample of m sequences of the form:

  Feature vectors: $o_1, o_2, \ldots, o_T \in \mathbb{R}^N$

  CD phones: $p_1, p_2, \ldots, p_l \sim ae_{c,t}$.

- Parameters:
  - mean and variance of Gaussians μ_j, σ_j.
  - mixture coefficients λ_k.
- Problems:
  - segmentation.
  - model initialization.

Estimation Algorithm

- Baum-Welsh algorithm:
  - maximum likelihood principle.
  - generalizes to continuous case with Gaussians and GMMs.
- Questions: segmentation, model initialization.

Univariate Gaussian - ML solution

- **Problem:** find most likely Gaussian distribution, given sequence of real-valued observations

$$3.18, 2.35, .95, 1.175, \ldots$$

- **Normal distribution:** $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$
- **Likelihood:** $l(p) = -\frac{1}{2} m \log(2\pi\sigma^2) - \sum_{i=1}^{m} \frac{(x_i - \mu)^2}{2\sigma^2}.$
- **Solution:** l is differentiable and concave;

$$\frac{\partial p(x)}{\partial \mu} = 0 \Leftrightarrow \mu = \frac{1}{m} \sum_{i=1}^{m} x_i \quad \frac{\partial p(x)}{\partial \sigma^2} = 0 \Leftrightarrow \sigma^2 = \frac{1}{m} \sum_{i=1}^{m} x_i^2 - \mu^2.$$

Gradients

- General identities:
  - log of determinant

$$\nabla_\sigma(\log det(\sigma)) = (\sigma^{-1})^\top = \sigma^{-\top}.$$

  - bilinear form

$$\nabla_\sigma(x^\top \sigma x) = x x^\top.$$

Multivariate Gaussian - ML solution

- **Log likelihood**: sample $x_1, \ldots, x_m$.

For each x_i, $\Pr[x_i] = \frac{1}{(2\pi)^{N/2}|\sigma|^{1/2}} \exp\Big(-\frac{1}{2}(x_i-\mu)^\top \sigma^{-1}(x_i-\mu)\Big)$.

$$L = \sum_{i=1}^m \log \Pr[x_i] = \sum_{i=1}^m -\frac{N}{2}\log(2\pi) + \frac{1}{2}\log|\sigma^{-1}| - \frac{1}{2}(x_i-\mu)^\top \sigma^{-1}(x_i-\mu).$$

- **ML solution**:

$$\frac{\partial L}{\partial \mu} = \sum_{i=1}^m \sigma^{-1}(x_i-\mu) = 0 \Rightarrow \boxed{\mu = \frac{1}{m}\sum_{i=1}^m x_i.}$$

$$\frac{\partial L}{\partial \sigma^{-1}} = \sum_{i=1}^m \frac{1}{2}(\sigma^\top - (x_i-\mu)(x_i-\mu)^\top) = 0 \Rightarrow \boxed{\sigma = \frac{1}{m}\sum_{i=1}^m (x_i-\mu)(x_i-\mu)^\top.}$$

GMMs - EM Algorithm

- Mixture of M Gaussians:

$$p_\theta[x] = \sum_{k=1}^{M} \lambda_k \mathcal{N}(x; \mu_k, \sigma_k) \qquad L = \sum_{i=1}^{m} \log \sum_{k=1}^{M} \lambda_k \mathcal{N}(x_i; \mu_k, \sigma_k).$$

- EM algorithm: **let** $p_{i,k}^t = \mathcal{N}(x_i; \mu_k^t, \sigma_k^t)$.
    - E-step: $$q_{i,k}^{t+1} = p_{\theta^t}[z = k | x_i] = \frac{\lambda_k^t p_{i,k}^t}{\sum_{k=1}^{M} \lambda_k^t p_{i,k}^t}.$$
    - M-step: $$\mu_k^{t+1} = \frac{\sum_{i=1}^{m} q_{i,k}^{t+1} x_i}{\sum_{i=1}^{m} q_{i,k}^{t+1}}$$

$$\sigma_k^{t+1} = \frac{\sum_{i=1}^{m} q_{i,k}^{t+1} (x_i - \mu_k^{t+1})(x_i - \mu_k^{t+1})^\top}{\sum_{i=1}^{m} q_{i,k}^{t+1}}$$

$$\lambda_k^{t+1} = \frac{1}{m} \sum_{i=1}^{m} q_{i,k}^{t+1}.$$

GMMs - EM Algorithm

- **Proof: M-step.**
  - Auxiliary function:

$$l = \sum_{k=1}^{M} \sum_{i=1}^{m} p_\theta[z = k | x_i] \log p_\theta[x_i, z = k] = \sum_{k=1}^{M} \sum_{i=1}^{m} q_{i,k} \log p_\theta[x_i, z = k].$$

$$= \sum_{k=1}^{M} \sum_{i=1}^{m} q_{i,k} \Big[\log \lambda_k - \frac{1}{2}(x_i - \mu_k^{t+1})^\top \sigma_k^{-1} (x_i - \mu_k^{t+1}) - \frac{1}{2} \log(2\pi) + \frac{1}{2} \log |\sigma_k^{-1}|\Big].$$

  - Optimization for fixed q and $\sum_{k=1}^{M} \lambda_k = 1$:

$$\frac{\partial L}{\partial \mu_k} = \sigma_k^{-1} \sum_{i=1}^{m} q_{i,k}(x_i - \mu_k) \quad \frac{\partial L}{\partial \lambda_k} = \frac{1}{\lambda_k} \sum_{i=1}^{m} q_{i,k} - \beta$$

$$\frac{\partial L}{\partial \sigma_k^{-1}} = \frac{1}{2} \sum_{i=1}^{m} q_{i,k}(\sigma_k^\top - (x_i - \mu)(x_i - \mu)^\top).$$

HMMs - EM Algorithm

- Use EM algorithm in discrete case to determine the probability of each transition e at time t : $w[e]$.
- Use GMMs update combined with the probability for each observation to be emitted from each transition e .

Initialization

- Selection of model:
  - HMM topology (states and transitions).
  - number of mixtures (M).
- Flat start: all distributions with same average values (mean and variance) computed over entire training set.
- Existing hand-labeled segmentation: partial segmentation basis.
- Uniform segmentation: equal number of HMM transitions per training example.

Forced Alignment

- Viterbi training: approximate but faster method to determine HMM path.
- Segmental K-means: approximate but faster method to determine which Gaussian of a mixture the training instance has been sampled from.

Viterbi Alignment

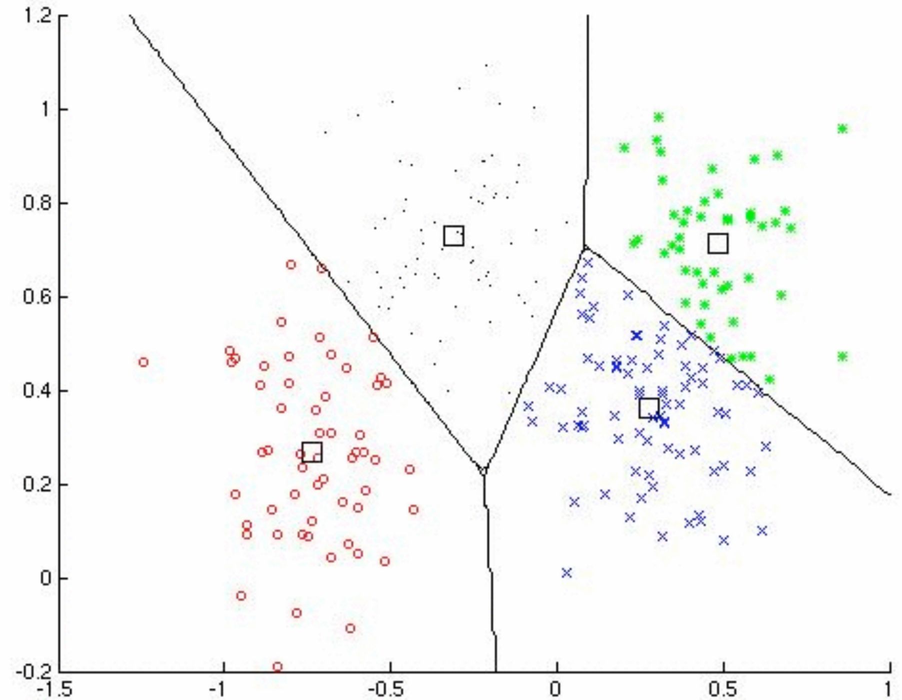
- **Idea**: faster alignment based on most likely path.
  - use current acoustic model.
  - align sequence of feature vectors with most likely path (best path algorithm, e.g., Viterbi).

o	1	2	3	4	5	6	7	8	9	10
e	e_{11}	e_{11}	e_{11}	e_{12}	e_{22}	e_{22}	e_{23}	e_{33}	e_{34}	e_{44}

Segmental *K*-Means

- Idea: use clustering algorithm to determine which Gaussian generated each observation.
- Solution: use *K*-means clustering algorithm to initialize distribution means.
  - Initialization: select *K* centroids $c_1, \ldots, c_K$.
  - Repeat until no centroid change:
    - for each point x_i find closest centroid c_j and assign x_i to cluster j.
    - for each cluster j, redefine c_j as the centre of mass.

Notes



- Convergence rate of K-means: subject of current research.
- GMM EM algorithm: soft version of K-means.

Variable Number of Mixtures

- Problems:
  - number of mixtures required.
  - possible over- or underfitting.
- Solution:
  - originally single Gaussian distribution.
  - create two-component mixture with slightly perturbed means $\mu \pm \epsilon$ with the same covariance matrix.
  - model parameters reestimated until desired complexity reached.

Acoustic Modeling

- In practice:
  - complicated recipes or heuristics with large number of *ad hoc* techniques.
  - key skilled human supervision: choice of initial parameters to avoid local minima, segmentation, choice and number of parameters.
  - computationally very costly: may take many days of several processors in large-vocabulary speech recognition.

Improvements

- Adaptation (VTLN, MLLR).
- Ensemble methods (ROVER).
- Better features (e.g., LDA, MMI).
- Discriminative training.

References

- L. E. Baum. An Inequality and Associated Maximization Technique in Statistical Estimation for Probalistic Functions of Markov Processes. *Inequalities*, 3:1-8, 1972.
- Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin . Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Vol. 39, No. 1. (1977), pp. 1-38..
- Jamshidian, M. and R. I. Jennrich: 1993, Conjugate Gradient Acceleration of the EM Algorithm. Journal of the American Statistical Association 88(421), 221-228.
- Lawrence Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of IEEE*, Vol. 77, No. 2, pp. 257, 1989.
- O'Sullivan. Alternating minimzation algorithms: From Blahut-Arimoto to expectation-maximization. Codes, Curves and Signals: Common Threads in Communications, A. Vardy, (editor), Kluwer, 1998.
- C. F. Jeff Wu. On the Convergence Properties of the EM Algorithm. The *Annals of Statistics*, Vol. 11, No. 1 (Mar., 1983), pp. 95-103.